### **Basic Statistics and Distributions**

### What is the difference between data analysis and machine learning?

### Data analysis requires strong knowledge of coding and basic knowledge of statistics.

### Machine learning, on the other hand, requires basic knowledge of coding and strong knowledge of statistics and business.

### 

### 2. What is big data?

### Big data has 5 major components – volume (size of data), velocity (inflow of data) and variety (types of data), veracity and value.

### Big data causes “overloads”

### 

### 3. What are the four main things we should know before studying data analysis?

### Descriptive statistics

### Inferential statistics

### Distributions (normal distribution / sampling distribution)

### Hypothesis testing

### 

### 4. What is the difference between inferential statistics and descriptive statistics?

### Descriptive statistics – provides exact and accurate information.

### Inferential statistics – provides information of a sample and we need inferential statistics to reach a conclusion about the population.

### 

### 5. What is the difference between population and sample in inferential statistics?

### From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

### From the sample we calculate the statistics

### From the sample statistics we conclude about the population

### 

### 6. What are descriptive statistics?

### Descriptive statistic is used to describe the data (data properties)

### 5-number summary is the most commonly used descriptive statistics

### 

### 7. Most common characteristics used in descriptive statistics?

### Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.

### Mean – average of all the numbers

### Median – the number in the middle

### Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.

### Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.

### Range = Max – Min

### InterQuartile Range (IQR) = Q3 – Q1

### Standard Deviation (σ) = √(∑(x-µ)2 / n)

### Variance = σ2

### Shape – the shape of the data can be symmetric or skewed

### Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

### Left skewed – the left tail is longer than the right side

### Right skewed – the right tail is longer than the left side

### Outlier – An outlier is an abnormal value

### Keep the outlier based on judgement

### Remove the outlier based on judgement

### 8. What is quantitative data and qualitative data?

### Quantitative data is also known as numeric data

### Qualitative data is also known as categorical data

### 

### 9. How to calculate range and interquartile range?

### IQR = Q3 – Q1

### Where, Q3 is the third quartile (75 percentile)

### Where, Q1 is the first quartile (25 percentile)

### 

### 10. Why do we need a 5-number summary?

### Low extreme (minimum)

### Lower quartile (Q1)

### Median

### Upper quartile (Q3)

### Upper extreme (maximum)

### 

### 11. What is the benefit of using box plots?

### Shows the 5-number summary pictorially

### Can be used to find outliers and compare group of histograms

### 

### 12. What is the meaning of standard deviation?

### It represents how far are the data points from the mean

### (σ) = √(∑(x-µ)2 / n)

### Variance is the square of standard deviation

### 

### 13. What is the left skewed distribution and right skewed distribution?

### Left skewed

### The left tail is longer than the right side

### Mean < median < mode

### Right skewed

### The right tail is longer than the right side

### Mode < median < mean

### 14. What does symmetric distribution mean?

### The part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

### Few examples are – uniform distribution, binomial distribution, normal distribution

### 

### 15. What is the relationship between mean and median in normal distribution?

### In the normal distribution mean is equal to median

### 

### 16. What does it mean by bell curve distribution and Gaussian distribution?

### Normal distribution is called bell curve distribution / Gaussian distribution

### It is called bell curve because it has the shape of a bell

### It is called Gaussian distribution as it is named after Carl Gauss

### 

### 17. How to convert normal distribution to standard normal distribution?

### Standardized normal distribution has mean = 0 and standard deviation = 1

### To convert normal distribution to standard normal distribution we can use the formula

### X (standardized) = (x-µ) / σ

### 

### 18. What is an outlier?

### An outlier is an abnormal value (It is at an abnormal distance from rest of the data points).

### 

### 19. Mention one method to find outliers?

### Shows the 5-number summary can be used to identify the outlier

### Widely used – Any data point that lies outside the 1.5 \* IQR

### Lower bound = Q1 – (1.5 \* IQR)

### Upper bound = Q3 + (1.5 \* IQR)

### 

### 20. What can I do with outliers?

Generally we first check the performance of model with outliers,Then we impute them and check the model performance.And even before that check first how much percent of data is outliers.Based on that you can take decision.

### Remove outlier

### When we know the data-point is wrong (negative age of a person)

### When we have lots of data

### We should provide two analyses. One with outliers and another without outliers.

### Keep outlier

### When there are lot of outliers (skewed data)

### When results are critical

### When outliers have meaning (fraud data)

21. What is the difference between population parameters and sample statistics?

Population parameters are:

Mean = µ

Standard deviation = σ

Sample statistics are:

Mean = x (bar)

Standard deviation = s

22. Why do we need sample statistics?

Population parameters are usually unknown hence we need sample statistics.

23. How to find the mean length of all fishes in the sea?

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

24. What are the effects of the width of the confidence interval?

Confidence interval is used for decision making

As the confidence level increases the width of the confidence interval also increases

As the width of the confidence interval increases, we tend to get useless information also.

Useless information – wide CI

High risk – narrow CI

25. Mention the relationship between standard error and margin of error?

As the standard error increases the margin of error also increases

26. Mention the relationship between confidence interval and margin of error?

As the confidence level increases the margin of error also increases

27. What is the proportion of confidence intervals that will not contain the population parameter?

Alpha is the portion of confidence interval that will not contain the population parameter

α = 1 – CL

28. What is the difference between 95% confidence level and 99% confidence level?

The confidence interval increases as me move from 95% confidence level to 99% confidence level

29. What do you mean by degree of freedom?

DF is defined as the number of options we have

DF is used with t-distribution and not with Z-distribution

For a series, DF = n-1 (where n is the number of observations in the series)

30. What do you think if DF is more than 30?

As DF increases the t-distribution reaches closer to the normal distribution

At low DF, we have fat tails

If DF > 30, then t-distribution is as good as normal distribution

31. When to use t distribution and when to use z distribution?

The following conditions must be satisfied to use Z-distribution

Do we know the population standard deviation?

Is the sample size > 30?

CI = x (bar) – Z\*σ/√n to x (bar) + Z\*σ/√n

Else we should use t-distribution

CI = x (bar) – t\*s/√n to x (bar) + t\*s/√n

32. What is H0 and H1? What is H0 and H1 for two-tail test?

H0 is known as null hypothesis. It is the normal case / default case.

For one tail test x <= µ

For two-tail test x = µ

H1 is known as alternate hypothesis. It is the other case.

For one tail test x > µ

For two-tail test x <> µ

33. What is p-value in hypothesis testing?

If the p-value is more than then critical value, then we fail to reject the H0

If p-value = 0.015 (critical value = 0.05) – strong evidence

If p-value = 0.055 (critical value = 0.05) – weak evidence

If the p-value is less than the critical value, then we reject the H0

If p-value = 0.055 (critical value = 0.05) – weak evidence

If p-value = 0.005 (critical value = 0.05) – strong evidence

34. How to calculate p-value using manual method?

Find H0 and H1

Find n, x(bar) and s

Find DF for t-distribution

Find the type of distribution – t or z distribution

Find t or z value (using the look-up table)

Compute the p-value to critical value

35. How to calculate p-value using EXCEL?

Go to Data tab

Click on Data Analysis

Select Descriptive Statistics

Choose the column

Select summary statistics and confidence level (0.95)

36. What do we mean by – making decision based on comparing p-value with significance level?

If the p-value is more than then critical value, then we fail to reject the H0

If the p-value is less than the critical value, then we reject the H0

37. What is the difference between one tail and two tail hypothesis testing?

2-tail test: Critical region is on both sides of the distribution

H0: x = µ

H1: x <> µ

1-tail test: Critical region is on one side of the distribution

H1: x <= µ

H1: x > µ

38. What do you think of the tail (one tail or two tail) if H0 is equal to one value only?

It is a two-tail test

39. What is the critical value in one tail or two-tail test?

Critical value in 1-tail = alpha

Critical value in 2-tail = alpha / 2

40. Why is the t-value same for 90% two tail and 95% one tail test?

P-value of 1-tail = P-value of 2-tail / 2

It is because in two tail there are 2 critical regions

41. **What is the central limit theorem?**

The central limit theorem (CLT) states that **the distribution of sample means approximates a normal distribution as the sample size gets larger**, regardless of the population's distribution

This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

42. **What is observational and experimental data in Statistics?**

Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

43. **What is meant by mean imputation for missing data?**

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

Mean imputation is not bad.It depends how much percentage of data is missing.If more than 50% is missing than mean imputation will not make sense as we will not have variance in data

### **44. What is an outlier? How can outliers be determined in a dataset?**

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

Outliers are determined by using two methods:

* Standard deviation/z-score
* Interquartile range (IQR)

### **45. How is missing data handled in statistics?**

There are many ways to handle missing data in Statistics:

* Prediction of the missing values
* Assignment of individual (unique) values
* Deletion of rows, which have the missing data
* Mean imputation or median imputation
* Using random forests, which support the missing values

46. **What are the types of selection bias in statistics?**

There are many types of selection bias as shown below:

* Observer selection
* Attrition
* Protopathic bias
* Time intervals
* Sampling bias

47. **What type of data does not have a log-normal distribution or a Gaussian distribution?**

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

48. **What is the meaning of the five-number summary in Statistics?**

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

* Low extreme (Min)
* First quartile (Q1)
* Median
* Upper quartile (Q3)
* High extreme (Max)

### **49. What is the meaning of standard deviation?**

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

### **50 What is a bell-curve distribution?**

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

### **51. What is skewness?**

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

### **52. What is kurtosis?**

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

### **53. What is correlation?**

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

### **54. What are left-skewed and right-skewed distributions?**

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

### **55. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?**

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

### **56.What is the relationship between the confidence level and the significance level in statistics?**

The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

Significance level = 1 − Confidence level

### **57.What is the relationship between mean and median in a normal distribution?**

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset’s mean and median.

### **58.What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?**

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

* The lower quartile (Q1) is the 25th percentile.
* The middle quartile (Q2), also called the median, is the 50th percentile.
* The upper quartile (Q3) is the 75th percentile.

### **59.How do the standard error and the margin of error relate?**

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

### **60. What is one sample t-test?**

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population’s mean.

### **61. What is an alternative hypothesis?**

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

### **62. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?**

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

### **63. What are the types of biases that we encounter while sampling?**

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis. There are three types of biases:

* The selection bias
* The survivorship bias
* The undercoverage bias

**64. What are some of the techniques to reduce underfitting and overfitting during model training?**

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

For reducing underfitting:

* Increase model complexity
* Increase the number of features
* Remove noise from the data
* Increase the number of training epochs

For reducing overfitting:

* Increase training data
* Stop early while training
* Lasso regularization
* Use random dropouts